Modularity and Extreme Edges of the Internet

Kasper Astrup Eriksen* and Ingve Simonsen†

The Nordic Institute for Theoretical Physics–NORDITA, Blegdamsvej 17, DK-2100 Copenhagen Ø, Denmark

Sergei Maslov‡

Department of Physics, Brookhaven National Laboratory, Upton, New York 11973

Kim Sneppen§

The Niels Bohr Institute, Blegdamsvej 17, DK-2100 Copenhagen Ø, Denmark
Department of Physics, The Norwegian University of Science and Technology, N-7491, Trondheim, Norway

We study the spectral properties of a diffusion process taking place on the Internet network focusing on the slowest decaying modes. These modes identify an underlying modular structure roughly corresponding to individual countries. For instance, in the slowest decaying mode the diffusion current flows from Russia to U.S. military sites. Quantitatively the modular structure manifests itself in a 10 times larger participation ratio of its slow decaying modes compared to a random scale-free network. We propose to use the fraction of nodes participating in slow decaying modes as a general measure of the modularity of a network. For the 100 slowest decaying modes of the Internet this fraction is ∼30%. Finally, we suggest that the degree of isolation of an individual module can be assessed by comparing its participation in different diffusion modes.

Virtually any complex system has an underlying network that defines the backbone of interactions among its components. Examples of such networks include the Internet and the World Wide Web, molecular networks of living cells, food webs in ecosystems, etc. An important question is whether nodes of such a network can be divided into smaller subnetworks (modules), which interact with each other relatively weakly [1]. Estimating the strength of intermodular interactions, localizing crucial links connecting these modules to each other, and finding pairs of modules which are the most distant from each other is important for several reasons. First of all, it serves as a test of stability of the system with respect to breaking it up into truly isolated components. Such a breakup would be undesirable in, for example, the Internet. Creation of extra connections between the most distant modules in the network and reinforcement of crucial links is an efficient way to increase its stability. Second, by measuring the relative strength of inter- and intramodular connections one directly assesses the quality of the independent module approximation, which may turn out to be important in modeling the actual dynamics of a given complex system.

In this work we explore the modular structure present in the physical layout of the Internet. To this end we study an auxiliary diffusion process taking place on this network. The slowest modes of diffusion, easily identifiable from the spectrum of its transfer matrix, allow us to detect the weakly interacting modules of the Internet. These modules turn out to roughly correspond to individual countries or for large countries to cultural or geographical regions within the country. Of course, the diffusion process studied in this work does not describe the real dynamics of the information flow over the Internet. However, the detected modular features play an important role in any local dynamical process taking place on this network including the real Internet traffic.

Analysis of spectral properties of a similar diffusion process lies at the heart of the popular search engine www.google.com [2]. Its variants have also been applied to social networks (the correspondence analysis) [3], random and small-world networks (the Laplace equation analysis) [4], artificial scale-free networks [5,6], and the community structure of the World Wide Web [7].

Here we study the physical backbone of the Internet on a coarse-grained level of the so-called autonomous systems (AS), which are large groups of routers and servers belonging to one organization such as a university or a business enterprise (e.g., an Internet service provider). For the information about which AS pairs directly exchange information with each other we use the 3 January 2000 data set of the National Laboratory for Applied Network Research. According to these data, the entire Internet back then consisted of 6474 autonomous systems exchanging information via 12 572 undirected links [8]. As expected for the Internet, any pair of autonomous systems is connected to each other by at least one path, so that topologically the network consists of just one large cluster. The diffusion process we analyze here describes the dynamics of a large number of random walkers moving on the network at discrete time steps. Statistical properties of returns to the origin of such random walks have recently been used to measure the effective dimensionality of several complex networks [9]. At each time

step every walker moves from its current node to one of the neighboring nodes along a randomly selected link. The average dynamics of this process is described by

$$\rho_i(t+1) = \sum_j T_{ij}\rho_j(t), \tag{1}$$

where $\rho_i(t)$ is the expectation value of the number of random walkers at site i and time t. The elements T_{ij} of the transfer matrix are equal to $1/K_j$ for neighboring nodes i and j and zero otherwise. Here K_j is the connectivity (the number of immediate neighbors) of the node j from which a walker steps to the node i. Note that $\sum_i T_{ij} = 1$, so that the total number of walkers is conserved at all times. Equation (1) can also be rewritten as a discrete time diffusion equation $\rho_i(t+1) - \rho_i(t) = \sum_j (T_{ij} - \delta_{ij})\rho_j(t)$. Hence the diffusion matrix $\mathbf{D}$ is related to the transfer matrix $\mathbf{T}$ simply as

$$\mathbf{D} = \mathbf{T} - \mathbf{1}. \tag{2}$$

As time advances the distribution of random walkers approaches a steady state $\rho_i(\infty)$ in which the diffusion current flowing from a node i to a node j is exactly balanced by that flowing from j to i. This is satisfied when the average number of walkers $\rho_i(\infty)$ on every node i is proportional to its connectivity K_i.

The relaxation of any initial distribution of random walkers among nodes, $\rho_i(0)$, towards the steady state configuration $\rho_i(\infty)$ is determined by the spectral properties of the matrix $\mathbf{T}$ [10] (or alternatively $\mathbf{D}$). For instance, the steady state configuration $\rho_i(\infty)$ itself is proportional to the principal eigenvector $\rho_i^{(1)}$ of $\mathbf{T}$ corresponding to its largest eigenvalue $\lambda^{(1)} = 1$, which is unique for single component networks such as the Internet. The remaining eigenvectors $\rho^{(\alpha)}$ describe the decay of the initial configuration towards the steady state with a characteristic decay time $\tau^{(\alpha)}$ related to the corresponding eigenvalue $\lambda^{(\alpha)}$ through $\exp(-1/\tau^{(\alpha)}) = |\lambda^{(\alpha)}|$. Note that in general there exist both nonoscillatory ($\lambda^{(\alpha)} \simeq 1$) and oscillatory ($\lambda^{(\alpha)} \simeq -1$) slowly decaying modes.

The modularity of a given complex network reflects itself in statistical properties of its diffusion eigenvectors $\rho_i^{(\alpha)}$. One such property is the participation ratio (PR), which quantifies the effective number of nodes participating in a given eigenvector with a significant weight. In the Internet the components of the principal eigenvector $\rho_i^{(1)} \propto \rho_i(\infty) \propto K_i$ as well as those of other slow decaying modes are broadly distributed (scale-free) [11] and as such tend to be localized on just a few highly connected nodes. In this case participation ratios are best calculated using the normalized eigenvector

$$c_i^{(\alpha)} = \rho_i^{(\alpha)}/K_i \tag{3}$$

of outgoing currents flowing from node i along each of its links. More formally $c_i^{(\alpha)}$ is also the eigenvector of the transposed transfer matrix $\mathbf{T}^\dagger$ with the same eigenvalue

$\lambda^{(\alpha)}$. For such a vector normalized by $\sum c_i^2 = 1$ the participation ratio is defined as $PR = (\sum_{i=1}^N c_i^4)^{-1}$.

In Fig. 1 the participation ratio of eigenvectors $c_i^{(\alpha)}$ (top) and the eigenvalue density (i.e., the spectrum of the matrix) (bottom) is plotted as a function of the corresponding eigenvalue $-1 < \lambda_i^{(\alpha)} < 1$. The data for the Internet (filled circles) is displayed together with the data for its randomized counterpart (open squares). The randomization of the Internet was performed in such a way that the connectivity of every node is strictly preserved [13]. It was argued [13,14] that such a network constitutes a proper null model of a given complex network. Since this random network has the same scale-free distribution of connectivities as the Internet [11] it will be referred to as a random scale-free network (RSFN).

Comparing the data for the Internet and an RSFN, we note that while the density of states is rather similar in these two networks [Fig. 1(b)], the participation ratio of the slowly decaying modes (especially for the nonoscillatory ones with λ close to 1) is markedly higher in the Internet than in an RSFN [Fig. 1(a)]. In these nonoscillatory modes the diffusion current flows from relatively isolated regions (modules) along the few links connecting them to the rest of the network. If for such a module these links would be hypothetically cut one by one, the corresponding eigenvalue would gradually increase towards unity, $\lambda^{(\alpha)} \to 1$, while the eigenvector would become more and more localized on the module. When finally the module is completely disconnected from the network, the eigenvector has evolved to the steady state solution on the module, which has the participation ratio equal to its size. Thus the PR of slowly decaying eigenmodes serves as a good quantitative estimate of the size
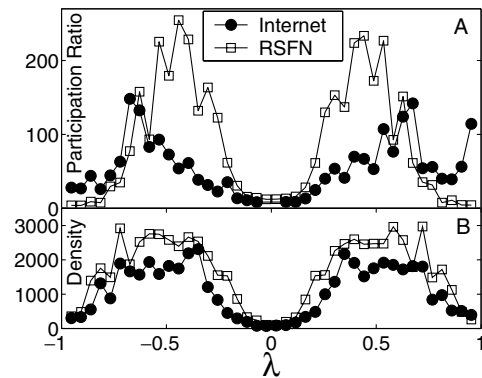

FIG. 1. The participation ratio $PR^{(\alpha)}$ (a) and the eigenvalue density (b) as a function of the eigenvalue $-1 < \lambda^{(\alpha)} < 1$ measured in the Internet (filled circles) and in its randomized counterpart (open squares)—a random scale-free network (RSFN). The participation ratio was averaged over λ bins of size 0.05 excluding eigenmodes $\lambda^{(\alpha)} = 0$ [12] and $\lambda^{(1)} = 1$. Notice that for $|\lambda| \simeq 1$ participation ratios in the Internet significantly exceed those in an RSFN indicating the modular character of the former network.

of modules in the network. In an RSFN these modules are small consisting of just a handful of nodes that accidentally happen to be loosely connected to the rest of the network. The fact that the participation ratios of slow decaying modes on the Internet significantly exceed those in an RSFN indicates that the corresponding modules are real and not accidental. The average participation ratio of slowly decaying modes can be quantified by $\sum_\alpha \mathrm{PR}^{(\alpha)}|\lambda^{(\alpha)}|^k / \sum_\alpha |\lambda^{(\alpha)}|^k$. For $5 \le k \le 10$ this average changes only slowly in both the Internet and an RSFN and equals approximately 60 and 5, respectively. A rough estimate for the number of different modules is given by the number of slowly decaying nonoscillatory states in Fig. 1(a) that have a participation ratio significantly exceeding that of an RSFN. For the Internet the number is around 100. The sum of the participation ratios for these first 100 modes, ~ 5400, providing a rough estimate of the total number of nodes in the modular part of the network, is about an order of magnitude larger than in an RSFN. If one takes special care to avoid double counting nodes that strongly participate in more than one of these slowly decaying eigenmodes, the number of nodes in the modular part of the network is reduced to ~ 1800. Thus the overall modularity of the Internet is at least $1800/6500 \simeq 30\%$.

To determine the organizing principle behind these Internet modules in Fig. 2 we plot the outgoing current $c_i^{(2)}$ in the slowest decaying diffusion mode ($\lambda^{(2)} = 0.9626$) as a function of the AS number (note that some AS numbers are not yet in use). Autonomous systems known to be located in Russia are marked with a circle. The PR for this eigenmode is 107, while the total number of Russian AS in our data set is 174. In Fig. 2 one can see that almost all the autonomous systems that significantly

participate in this mode (large positive $c_i^{(2)}$) are Russian. We have checked that the few exceptions to this rule are, in fact, autonomous systems closely related to Russia. Thus in the slowest decaying mode the diffusion current flows from a module that may be identified with Russia towards the rest of the Internet. Curiously enough the set of autonomous systems farthest away from Russia (the most negative $c_i^{(2)}$) are located in the U.S. and belong to the U.S. military. This possible legacy of the cold war makes Russia and the U.S. military the extreme edges of the Internet. Performing a similar analysis for other slowly decaying modes, we get a similar picture, just with other pairs of countries being pulled out. For the Internet the modules thus correspond to individual countries, or for large countries to organizational or geographical features within the country.

It is interesting to note that these country modules cannot be detected using the spectral analysis of the adjacency matrix of the network [5,6,15]. The elements of this matrix, closely related to **T**, are equal to 1 for a pair of neighboring nodes and 0 otherwise. The largest eigenvectors of the adjacency matrix are known to be localized primarily on the highest connected hubs and their neighbors [5,6]. However, unlike in the case of **T**, this undesirable localization cannot be properly eliminated simply by dividing the components of the eigenvectors by the connectivity K_i. Hence eigenvectors of the adjacency matrix do not properly reflect the country-based modular structure uncovered in this work.

Having established that the Internet is indeed modular, we now address the question of how good these individual modules are. To this end we compare different eigenmodes $c_i^{(\alpha)}$ to each other. Although the primary feature in a slowly decaying eigenmode $c_i^{(\alpha)}$ is the flow between a dominant pair of country modules (such as between Russia and the U.S. military in $c_i^{(2)}$), other modules may also participate in it but to a smaller extent. This gives rise to a fine structure within slow decaying modes that is not captured by the participation ratio. The hallmark of a good module is that even though it participates in different eigenmodes α to a different extent, the relative distribution of α currents within the module stays approximately the same. In other words, it enters different eigenmodes as just 1 degree of freedom. In this case the ratio $c_i^{(\alpha)}/c_j^{(\alpha)}$ is approximately independent of the eigenmode α for any pair of nodes i and j within the module. This is equivalent to the condition that for any two different eigenmodes α and β, $c_i^{(\alpha)}/c_i^{(\beta)} = \mathrm{const}$ for every node i belonging to the given module. In Fig. 3 we plot the outgoing currents in the two slowest decaying nonoscillatory eigenmodes—$c_i^{(2)}$ and $c_i^{(3)}$—as a function of each other. Similar plots can be made for other pairs of slow decaying eigenmodes. The principal feature in this kind of plot is a starlike shape, where different rays of the star correspond to individual country modules. This type of plot is more powerful in identifying
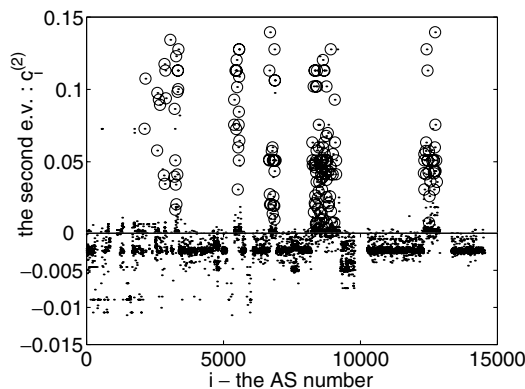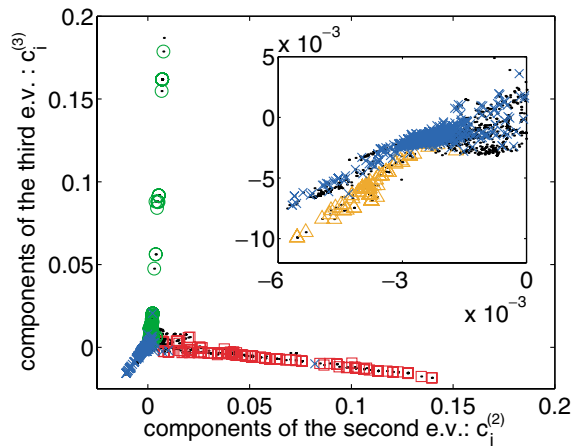
FIG. 2. Components $c_i^{(2)}$ of the slowest decaying diffusion mode in the Internet (eigenvalue $\lambda^{(2)} = 0.9626$) as a function of the AS number. The AS known to be geographically located in Russia are marked with circles. The scale of the negative part of the y axis is increased for clarity. Out of 100 autonomous systems with the most negative components $c_i^{(2)}$, those 23 for which we were able to find the description are associated with the U.S. military.

FIG. 3 (color). The Internet clustering: Coordinates of the ith AS in this plot are its components $(c_i^{(2)}, c_i^{(3)})$ in the two slowest decaying nonoscillatory diffusion modes. The color code reveals the geographical location of the AS (Russia: red squares; France: green circles; U.S.: blue crosses; Korea: orange triangles). Note the straight lines corresponding to good country modules.

individual modules than the participation ratio alone. Indeed, in Fig. 3 one can easily detect not only the most excited modules like Russia, France, and U.S. (red squares, green circles, and blue crosses), but also less excited ones like Korea (orange triangles). We believe that the idea of measuring the quality of individual modules by how proportionally their nodes participate in different slowly decaying modes can be easily generalized to other dynamical processes taking place on the network such as, e.g., spin dynamics, vibrational modes, etc.

In summary, we have demonstrated how a diffusion process taking place on the Internet network allows one to extract information about its modules and extreme edges. For many "real-world" complex networks the local context of a node (in terms of the linkage pattern) reflects or, perhaps, even determines the importance and function of the given node. For instance, in biology one can successfully assign putative functions to unclassified proteins based on the function of their interaction partners [16]. In general, the diffusion process introduced in this work can be seen as a systematic way to explore the local linkage structure of a network beyond just the nearest neighbors. The detection of the modular structure of a network is just one possible application of such a process [2,9].

*Electronic address: kasper@nordita.dk
†Electronic address: ingves@nordita.dk
‡Electronic address: maslov@bnl.gov
§Electronic address: sneppen@nbi.dk

[1] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821–7826 (2002).
[2] S. Brin and L. Page, in *Proceedings of the 7th International World Wide Web Conference* (Elsevier, Amsterdam, 1998).
[3] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications* (Cambridge University Press, Cambridge, 1994).
[4] R. Monasson, Eur. Phys. J. B **12**, 555 (1999).
[5] I. J. Farkas, I. Derényi, A.-L. Barabasi, and T. Vicsek, Phys. Rev. E **64**, 026704 (2001).
[6] K.-I. Goh, B. Kahng, and D. Kim, Phys. Rev. E **64**, 051903 (2001).
[7] J. Kleinberg, in *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms, 1998* (Association for Computing Machinery, New York, 1998); D. Gibson, J. Kleinberg, and P. Raghavan, in *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia, 1998* (ACM Press, New York, 1998).
[8] The data set collected by the Oregon Views project was downloaded from http://moat.nlanr.net/AS/.
[9] S. Bilke and C. Peterson, Phys. Rev. E **64**, 036106 (2001).
[10] Even though T_{ij} is not symmetric, all of its eigenvalues and eigenvectors are real, since it is related by a similarity transformation to the symmetric matrix $\tilde{T}_{ij} = T_{ij}\sqrt{K_j/K_i}$. $\tilde{T}_{ij}$ has exactly the same set of eigenvalues $\lambda^{(\alpha)}$ as T_{ij}, and the corresponding eigenvectors $\tilde{\rho}_i^{(\alpha)}$ are related to the eigenvectors $\rho_i^{(\alpha)}$ of T_{ij} via $\tilde{\rho}_i^{(\alpha)} = \rho_i^{(\alpha)}/\sqrt{K_i}$.
[11] M. Faloustos, P. Faloustos, and C. Faloustos, in *Proceedings of the ACM SIGCOMM '99 Conference, Cambridge, MA* (Association for Computing Machinery, New York, 1999), p. 251.
[12] Highly degenerate zero modes ($\lambda^{(\alpha)} = 0$) of a matrix do not have a well defined PR. These modes are, in fact, rather abundant and constitute $\approx 66\%$ of all the eigenvalues in the Internet and $\approx 50\%$ in an RSFN. Zero modes come in part from pairs of nodes of connectivity 1 interacting with the same highly connected partner. This structure is known to be overrepresented in the real Internet compared to an RSFN [13]. More generally zero modes are localized on small substructures in the network where every second node is of low connectivity.
[13] S. Maslov, K. Sneppen, and A. Zaliznyak, cond-mat/0205379; See also the chapter by S. Maslov, K. Sneppen, and U. Alon, in the *Handbook on Graphs and Networks*, edited by S. Bornholdt and H.-G. Schuster (Wiley-VCH Publishers, New York, 2002).
[14] S. Maslov and K. Sneppen, Science **296**, 910 (2002).
[15] D. M. Cvetkovic, M. Doob, and H. Sachs, *Spectra of Graphs* (Academic Press, New York, 1979).
[16] B. Schwikowski, P. Uetz, and S. Fields, Nat. Biotechnol. **18**, 1257–1261 (2000); S. Oliver, Nature (London) **403**, 601 (2000).